Artificial Intelligence Tools and Open Data Practices for EPA Chemical Hazard Assessments

Proceedings of a Workshop—in Brief

The U.S. Environmental Protection Agency's (EPA's) Integrated Risk Information System (IRIS) Program identifies and characterizes the human health hazards of chemicals found in the environment. Human health risk assessments cover hazard identification as well as dose-response analyses for cancer and noncancer outcomes that are obtained from IRIS assessments. Human health risk assessments are highly important as they are used to inform a broad range of risk-related decisions across the agency. These assessments involve systematic reviews (SRs) of the scientific literature, which obtain, evaluate, and summarize information to answer a research question in a transparent manner.

The Center for Public Health and Environmental Assessment (CPHEA) within EPA's Office of Research and Development (ORD) requested that the National Academies of Sciences, Engineering, and Medicine convene a workshop (see Box 1) to explore opportunities and challenges in using advances in artificial intelligence (AI) and data science to enhance human health risk assessments. The workshop was held virtually on May 25 and 26, 2022.

SR methods use rigorous, pre-established protocols in which trained professionals strive to find all relevant studies, extract information concerning the reported methods and findings, critically analyze the information, and summarize the information in a reusable manner. Because SRs for human health questions can involve consideration of a vast number of studies, they tend to be labor intensive and costly. These constraints apply not only to EPA but also to other government and nongovernmental organizations in the United States and abroad that conduct SRs to address environmental health questions.

Although various software is now commonly used to help screen studies for inclusion in SRs, the process is still time intensive. Additionally, subsequent steps, such as the extraction of data from the literature, continue to be done manually.

Recent advances in AI, machine learning, and data science hold the promise of using computer-assisted tools to increase the efficiency of SR methods. These tools could enable more rapid screening of the literature and may allow for automated data extraction. However, potential concerns about the reliability and reproducibility of the

BOX 1
Workshop Presenters and Discussants

Day 1

Topic: Promises and Prospects of AI and Data Science Applications
Moderator: Chirag Patel, Harvard Medical School
- Daniel Ho, Stanford University
- Christopher Mungall, Lawrence Berkeley National Laboratory
- Nicole Kleinstreuer, National Institute of Environmental Health Sciences (NIEHS)

Topic: Challenges for Applying AI to SR Methods
Moderator: Scott Auerbach, NIEHS
- Malcolm Macleod, University of Edinburgh
- Rens van de Schoot, Utrecht University
- Karen A. Robinson, Johns Hopkins University
- Olwenn Martin, University College London

Topic: Optimizing Data Extraction for Evidence Synthesis
Moderator: Byron Wallace, Northeastern University
- Weida Tong, U.S. Food and Drug Administration
- Jason Fries, Stanford University
- Daniel Sanders, IBM
- Marianthi-Anna Kioumourtzoglou, Columbia University

Day 2

Topic: AI Tools and Resources in SR
Moderator: David Reif, North Carolina State University
- Andrew Rooney, NIEHS
- Vickie Walker, NIEHS
- Brian Howard, Sciome, LLC
- Nancy Baker, Leidos
- Karen Ross, Georgetown University
- Mark Musen, Stanford University
- Julie McMurry, University of Colorado School of Medicine

Topic: SR Tools
Moderator: Joyce Tsuji, Exponent, Inc.
- Ryan Jones, U.S. Environmental Protection Agency (EPA)
- Sean Watford, EPA
- Derek Lord, Evidence Partners
- Eitan Agai, PICO Portal
- Artur Nowak, Evidence Prime
- Iain Marshall, King's College London

Topic: Ensuring Rigor and Reproducibility
Moderator: David Reif, North Carolina State University
- Marzyeh Ghassemi, Massachusetts Institute of Technology
- Chirag Patel, Harvard Medical School
- John Absher, Squarespace, Inc.

results from the application of these tools, as well as limitations of natural language processing (NLP), could complicate efforts to apply the tools more widely.

In setting the stage for the workshop, Kristina Thayer (Chemical and Pollutant Assessment Division, CPHEA/ORD, EPA) indicated that over the past several years there has been substantial onboarding of machine learning applications to help with the process of screening studies for further consideration. There is also interest in using AI tools to facilitate the process of extracting data on study design and results from the scientific literature.

Workshop presentations and panel discussions are summarized in this document. Posters presented during the workshop are available on the National Academies website.[1]

[1] See https://www.nationalacademies.org/event/05-25-2022/workshops-to-support-epas-development-of-human-health-assessments-artificial-intelligence-and-open-data-practices-in-chemical-hazard-assessment.

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE APPLICATIONS: PROMISES AND PROSPECTS

Daniel Ho (Stanford University) began the session by providing an overview of the use of innovative AI tools in the federal government. He noted that 45% of federal agencies are experimenting with AI but that federal officials have faced challenges of sophistication, accountability, and explainability. As an example, he cited the inability to identify the source of errors in the results of applying biometric scanning technology. He also noted capacity issues, including the importance of devoting adequate human capital to a project. He highlighted blended expertise, both technical knowledge of AI and the ability to draw on recent advances as well as subject-matter expertise relevant to the problems being solved.

Christopher Mungall (Lawrence Berkeley National Laboratory) highlighted the challenge that arises when the knowledge landscape is fragmented, existing as natural language text in the literature. He discussed how ontologies (formally defined vocabularies) can help to organize and annotate the information for the application of AI tools. Mungall described the gene ontology project as an example of the use of ontology for gene function.

He also described an upsurge of interest in the use of knowledge graphs for integrating data and applying machine learning in the biosciences. However, the knowledge graphs are not compatible with one another. To address this, Mungall is working on a biomedical translator project with the National Center for Advancing Translational Sciences at the National Institutes of Health to develop a standard data model, referred to as Biolink Model, for knowledge graphs. The project seeks to bring together a number of different automated systems, knowledge sources, and knowledge providers, so that a user can ask questions such as what chemicals or drugs may be used to treat certain neurological disorders that are associated with particular gene variants.

Nicole Kleinstreuer (National Toxicology Program [NTP]/ National Institute of Environmental Health [NIEHS]) proposed to define AI as "augmented intelligence" rather than "artificial intelligence" to avoid the notion that the goal of AI is to replace human intelligence. The concept of augmented intelligence, she said, encompasses the use of data science and computational tools to enhance and support the human intellect in generating insights into human disease processes and their susceptibility to environmental perturbation.

Kleinstreuer discussed the use of machine learning approaches to train quantitative structure activity relationship (QSAR) models. QSAR models are used to find complex relationships between chemical features, such as molecular structures, physicochemical properties, and toxicity values. She indicated that the NTP open-access tool OPERA (Open Quantitative Structure-activity/ property Relationship App or OPEn (q)saR App) that is a free resource that provides toxicity predictions based on chemical structures. The downloadable, open-access modeling suite predicts the toxicity of agents, such as flame retardants, pesticides, and air pollutants, revealing what factors are likely to contribute to or worsen adverse health outcomes. OPERA contains predictions for nearly 1 million chemical structures, training videos, testing and assessment strategies, computational models, and workflows to analyze chemical data.

Kleinstreuer highlighted the need for high-quality, openly available, well-annotated datasets; computing resources; and data infrastructure to foster the use of AI tools for addressing environmental health questions.

Chirag Patel (Harvard Medical School) served as the moderator for a panel discussion among the presenters. The topics discussed included how to weigh the evidence to be used for taxonomizing; recruiting individuals with the expertise needed to develop and apply AI methods; obtaining training data for AI tools, transparency in the use of complex models; and various uses of crowdsourcing approaches. Patel also asked the panelists for their opinions about transformative opportunities over the next few years.

Mungall mentioned hybrid systems that combine deep learning with older AI methods and better enforcement of standards and good data practice for making data findable, accessible, interoperable, and reusable.

Kleinstreuer listed increased dialogue with end users of AI tools, regulatory decision-makers, and the interested public. She also mentioned systems models, citing genome-scale metabolic models and agent-based modeling as examples. Another promising area is the development of compute-optimal large language models, she said, such as the DeepMind model named Chinchilla, which significantly outperforms the existing models.

Ho mentioned advances in neuromorphic computing that may radically improve the capabilities of ordinary researchers, the social gains possible from the application of AI tools, and new kinds of government–academies partnerships for better information exchange concerning AI tools development and applications. He noted the promising potential for a national AI research resource.

ADDRESSING CHALLENGES FOR APPLYING SYSTEMATIC REVIEW METHODS USING ARTIFICIAL INTELLIGENCE

In the next session of the workshop, Malcolm Macleod (University of Edinburgh) discussed the limitations of human screening for SRs that can become incorporated into computer-assisted screening tools. Macleod stated that the gold-standard approach, defined as using two human reviewers to perform title and abstract screening, is in need of improvement. He said that approach can miss many of the relevant results, especially when multiple experiments are included in one publication. Macleod discussed the development of tools to address this challenge by providing automatic PICO (population, intervention, comparison, and outcomes) extraction from abstracts and semi-automated data extraction via graphs, which save time without sacrificing accuracy.

Macleod raised a question concerning the amount of data loss that should be considered acceptable. He noted that only a limited amount of work has been done to assess the impact that failing to identify information sources has on SR conclusions. However, Macleod added, it is clear that a certain amount of missing data would not make a material impact on the conclusions one would draw.

Macleod indicated that he imagines a central data store in which an annotated label on a citation source from one SR is suitable and accessible for reuse by others. He added that this could be supplemented by the automated annotation of relevant literature.

The problem of data overwhelm was described by Rens van de Schoot (Utrecht University). When a person searches a database to answer a question, two problems come up, he said: (1) there are too many papers to read and (2) there are too few relevant papers. With his open-source project, ASReview, van de Schoot focuses on active learning, which is machine learning that can be deployed to present the most relevant paper, he said. Though there are differences across datasets and machine learning models, he said that "all of the simulation studies show that active learning outperforms human reading by far."

van de Schoot presented several principles in using any software package that implements AI:

- humans being in control,
- a completely open and transparent application,
- application using an unbiased estimate,
- being aware of when an AI-aided interface is used, and
- garbage in, garbage out.

Scott Auerbach (NIEHS) moderated a panel discussion among the two presenters, Karen A. Robinson (Johns Hopkins University School of Medicine) and Olwenn Martin (University College London).

Robinson noted that a fundamental challenge is the lack of standards that govern SRs. Martin stated that in chemical risk assessment, researchers might struggle with identifying chemicals properly, depending on how they are labeled in the literature. With toxicology, the guideline studies are relatively uniform in their reporting, Auerbach added, but traditional manuscripts remain non-standardized.

Martin expressed hesitation about fully automated data extraction. van de Schoot asked about trust in AI such that a researcher could set rules and let the machine do the work. Robinson said she believes that all parts of SR could be assisted by AI tools. van de Schoot underscored the aspect of AI assistance rather than full automation.

Macleod offered what he characterized as a practical response. “I think that having some measure of the provenance of the claim that you’re making as a regulator is important,” he said, adding that this approach asks the regulator to acknowledge the percentage of data (typically up to 10%) that is missing from the analysis.

OPTIMIZING DATA EXTRACTION FOR EVIDENCE SYNTHESIS AND HIGH-LEVEL DECISION-MAKING

Weida Tong (U.S. Food and Drug Administration [FDA]) discussed AI4TOX, an FDA program that focuses on applying new AI methods in the field of toxicology to inform regulatory decision-making. Tong explained that as part of the program they are developing a model called AnimalGAN that will use generative adversarial networks (GANs) to learn from past animal studies in such a way that it can generate animal study results for new and untested compounds without conducting new in vivo animal studies. They developed another model called Tox-GAN to determine the underlying mechanism of the toxicity using gene expression data. He indicated that AnimalGAN and Tox-GAN can be extremely helpful in generating toxicological information on chemicals on the basis of toxicological information on other chemicals.

Jason Fries (Stanford University) discussed programmatic labeling for data-centric NLP. He indicated that for traditional supervised learning, experts observe and label examples from some data distribution and use those labels to train some sort of model. This method is expensive and slow, and it is difficult to change or revisit decisions that were made for generating that data. For the past several years, he has been exploring the use of programmatic labeling or weak supervision. Instead of having domain experts manually label single data points, the focus is on designing labelers as a general concept, considering aspects such as rules and interactions with knowledge bases or ontologies. The objective is to automatically generate a training label on a fundamentally unlabeled set of data, which are then used to train a machine learning model.

According to Fries, programmatic labeling is a consistent and reproducible approach. It breaks down a problem, which an expert might encode in a single label, into a system of modular components. That offers opportunities to audit training data and interrogate assumptions used to generate training data.

Fries said a potential direction for programmatic labeling is natural language prompting in which people write insights as natural language instruction, instead of using programming language. This could open the labeling process to be more accessible to domain experts who are not trained in writing code.

Daniel Sanders (IBM) spoke about IBM’s use of AI tools to help in discovering safer chemicals and materials that are used in the production of computer chips. It involves incorporating chemical hazard assessments into AI-based design workflows for new hypothetical chemicals. Sanders provided a use case example of photo acid generators, which are a class of molecules used to pattern semiconductor devices. To develop alternatives with better environmental health and safety characteristics, Sanders’s team trained AI models to produce tens of thousands of chemical candidates, then screened them for further evaluation. He indicated that as AI models get more sophisticated, there is a need for a platform that allows multiple domain experts to interact with AI specialists.

Sanders said there is a need for improved data standards for publication and patenting such that existing and future NLP tools can more readily integrate information.

Marianthi-Anna Kioumourtzoglou (Columbia University) discussed how machine learning can assist epidemiological studies of chemical mixtures. A well-defined research question is crucial, she said, because it enables the identification of the best methods for finding answers—and to then state this question explicitly in the paper so an AI tool can readily capture that information and classify the paper for SR.

Kioumourtzoglou also discussed random sampling in machine learning and the effect of seed selection. All of the methods she examined exhibited some seed-dependent variability in the results. She indicated that the degree of variability differed across methods, the kind of organic congeners examined, and would likely vary with sample size. Kioumourtzoglou stated that seed sensitivity analysis can help evaluate robustness and interpretability. However, if the results are highly variable across seeds, it would help researchers extract information for SRs if the distributions of the estimated effects across seeds were presented.

In a panel moderated by Byron Wallace (Northeastern University) participants discussed various topics, including practical limitations and needs related to building and deploying AI solutions. Fries and Wallace noted an institutional issue of making a case to justify the financial resources needed to obtain, apply, and maintain AI systems.

Sanders indicated that data access is a barrier because of the reluctance to share data and analysis plans. That makes industry-wide collaboration difficult. He pointed to the need for advancements in federated learning and encryption, and to find a way for AI models to learn in the absence of data sharing. Kioumourtzoglou said that training people can be an issue, and Tong added that reliability, interpretability, applicability, and reproducibility are also pressing concerns.

USING ARTIFICIAL INTELLIGENCE TOOLS AND RESOURCES IN SYSTEMATIC REVIEW

The second day of the workshop covered how machine learning tools are being used—and may be used in the future—with a focus on SRs. Andrew Rooney (NTP/NIEHS) opened the day with an overview. He noted that a key challenge for SRs of environmental health hazards is their broad scope. For instance, all health effects for an individual chemical may be examined in a series of SRs. Other environmental questions examined through SR concern all of the exposures that might be associated with a particular disease or health effect, mixtures, or chemical classes (e.g., the organohalogen flame retardants,[2] the per- and polyfluoroalkyl substances, etc.). The SRs consider multiple evidence streams including human and animal studies as well as a diversity of other studies comprising pharmacokinetics, mechanistic evidence, and exposure information. Additionally, the relevant data are published in diverse forms and the studies are heterogeneous in design. He noted the diversity of study types and endpoints relevant for all evidence streams. This complexity presents a challenge for identifying and extracting data as well as for developing models, he said. Rooney also highlighted the challenge of annotating studies. He noted the need for annotated datasets "from a really diverse spectrum, so that the models can be applied to the diverse and heterogeneous data we need to make our decisions."

The next section of the workshop moderated by David Reif (North Carolina State University) provided an overview of a set of tool demonstrations that were pre-recorded.

Vickie Walker (NTP/NIEHS) presented Dextr, a tool that supports semi-automated data extraction using machine learning models. Dextr processes full-text PDFs and enables researchers to share data in a machine-readable format that can be used by model developers. Users can upload several PDFs at one time, import references, and use a data-cleaning module. The tool currently focuses on extracting metadata to support evidence maps, but future plans call for the extraction of study results, Walker said.

Brian Howard (Sciome, LLC) presented a platform called Swift AI, which combines three tools: Swift Review (evidence mapping, exploratory analysis), Active Screener (prioritizing studies), and FIDDLE 2.0 (extracting data). Swift Review and Active Screener focus on describing the problem and choosing the studies that will work best with machine learning. FIDDLE 2.0 extracts text from PDFs. Howard noted that this is not straightforward because the PDF is an image-driven format and was not developed to preserve the underlying text flow. FIDDLE 2.0 uses an algorithm to retain word order and placement across columns and tables. Overall these tools aim to introduce efficiencies, such as by pre-populating forms,

[2] See https://pubmed.ncbi.nlm.nih.gov/31436945.

and making inferences about the user's intent based on interaction and context, Howard said.

An Excel-based tool called Abstract Sifter was presented by Nancy Baker (Leidos). Abstract Sifter enhances the search capabilities of PubMed by making searches effective, triaging results, and tracking articles of interest. This tool also provides an overview of the literature for a set of chemicals or genes. The next version (7.1) will be enabled with an application programming interface to allow the user to obtain the DSSTox[3] ID for a list of chemicals in a PDF table or from medical subject headings (MeSH) terms. With the DSSTox ID, the chemical structure, CAS number, and toxicology profile, including whether the chemical has been tested, are all available in one tool, Baker said.

Karen Ross (Georgetown University) discussed her work with UniProt, a hub of protein sequence and function data. She said that a major challenge is the annotation of protein sequences. "We believe that the solution is to enhance our ability to automatically extract protein information from the scientific literature," Ross said, as well as a tool that would "automatically predict annotations for poorly understood proteins using information from the manually curated entries." Her team has been seeking out text-mining ontologies and functional prediction tools, and it has developed an ontology project called PRO, which is the reference ontology for proteins and protein forms in the Open Biomedical and Biomedical Ontologies Foundry. Because PRO is distributed in standard formats like OWL, it can be searched with the database query language SPARQL. This "makes it easier to integrate PRO information with information from other ontologies on the semantic web."

Mark Musen (Stanford University) described a system called BioPortal, which is an open repository for biomedical ontologies. He described Annotator, a tool that allows users to "relate journal articles to standardized terms in a way which allows us to cluster articles [and] to count articles." This offers users the ability to identify the ontologies that might work well with the text they are annotating. Researchers can also save their own ontologies. Musen also shared details about the Center for Expanded Data Annotation and Retrieval (CEDAR), which aims to "create standards-compliant metadata that would enhance the fairness of datasets that experimenters are putting into online repositories." With CEDAR, researchers can write a template that will follow a standard reporting guideline for discussing a given domain of science. Musen said that this helps users feel "confident about the metadata, and the fact that the metadata are saying what they need to say about experiments, so that those third parties can actually understand what was done in those various investigations."

Panel discussant Julie McMurry (University of Colorado School of Medicine) began with a few notes about data for decision-making. She emphasized that it is vital for AI tools to be assistive but not prescriptive, with a feedback loop to steer the AI in the right direction to improve the algorithm itself. Musen noted that investment in high-quality metadata is essential and will enhance understanding of what the datasets represent and how the experiments were done. In the future, there might be a big cultural shift, she added, whereby researchers will view "the primary output of the search as being the data, rather than the publication."

The panel touched on the issues of science literacy, public engagement, and metadata as well as crowdsourcing. Rooney noted that a focus on reporting objective measures is key to increasing public confidence, as quality judgments and synthesis decisions can be sources of disagreement. New ways of crediting contributions to publications could incentivize the availability of metadata, noted Reif. Community curation is possible and UniProt's interface allows anyone to suggest a relevant paper for a protein entry and also propose annotations for the protein from that paper, Ross said. "It's our experiment with trying to get more crowdsourcing of protein annotations."

[3] See https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database.

Elaine Faustman (University of Washington) noted that the World Data System's Scientific Committee (which she co-chaired), a subsection of the International Science Council, endorsed the idea of professional code developers receiving more credit as a means to encourage future developments in this area.

SYSTEMATIC REVIEW TOOLS

Ryan Jones (EPA) introduced a database system resource called HERO (Health and Environmental Research Online). The system was developed to support EPA's integrated science assessments, which have grown substantially over the past three decades. HERO is a repository of citations that can be screened, categorized, and shared. "If something gets cited, we have a copy of it," Jones said, which enables the agency to be transparent with the public about the cited literature. HERO's screening tools enable search results to be viewed by field of expertise, which is efficient for the multidisciplinary teams processing literature searches. HERO also offers a keyword-free search called citation mapping, which Jones said has been demonstrated to be "three or four more times likely to return results that we actually need than the traditional keyword search." In addition, Jones said HERO supports third-party tools such as Distiller and Swift.

Sean Watford (EPA) demonstrated an SR tool called Health Assessment Workspace Collaborative (HAWC) that was developed by EPA's Andy Shapiro. HAWC is used in assessment programs across EPA and tackles interoperability, which Watford described as a major challenge. Watford shared an example of using the Environmental Health Vocabulary, a controlled vocabulary that applies to animal health outcomes, but that may be expanded in the future to try to bring more structure to the extracted data. This opens the opportunity to reuse data, including as a training dataset for model development.

The next tool, DistillerSR, was presented by Derek Lord (Evidence Partners). A Software-as-a-Service-based platform, DistillerSR's objective is to help make literature reviews faster and easier, Lord said, throughout the entire review lifecycle, beginning with a search. DistillerSR tracks and manages reviews and published references in one place. Users can also "extend DistillerSR into different data platforms, such as safety databases, predictive analytics, or AI-based tools," he said. Lord shared an example use case of DistillerSR's intelligent workflows and AI to stay updated with COVID-19 data. "They were able to streamline the process using AI classifiers to help screen and complete the review," Lord said. DistillerSR cut the literature review screening time in half. DistillerSR is able to de-duplicate the results of searches, check for errors during the screening process, and also support the data extraction component, classifying and organizing it with premade templates, and, finally, reporting that data and pulling it out of the system.

Eitan Agai (PICO Portal) shared his work using automation (machine learning and NLP algorithms) to expedite SRs. PICO Portal, an online SR platform, is designed to serve evidence synthesis projects of any size or scale, Agai said. Agai placed PICO Portal in the company of other SR software programs, including abstrackr, DistillerSR, and Rayyan. Agai shared how his prior experience in the mortgage and financial industry translated to his current project. "I found out that systematic review has similar pain points as mortgage automation," he said, outlining the process of classifying documents, extracting data from documents (bank statements, paystubs, W-2s, etc.), and following guidelines and regulation. PICO Portal uses machine learning to help classify documents and prioritize articles during the screening process and provides analytics such as user accuracy and speed. His team received an NIEHS grant to continue the work of pairing NLP with ontology.

The next tool was Laser AI, demonstrated by Artur Nowak (Evidence Prime), who indicated that the company's approach is largely one of augmented intelligence, an idea Kleinstreuer introduced earlier. Rather than trying to replace human decisions and human intelligence, they are trying to enhance them using AI tools, Nowak said. "I hope that this mix of human and artificial intelligence will save us from the information overload," he added. Nowak shared case studies from the clinical space, including an SR for the

World Health Organization in which Laser AI acted as a third reviewer. The role of the SR tool is to connect publications, entities, and ontologies. "As AI, we provide some sort of a link that is then validated by people as part of their systematic review work," he said. He also indicated that it would be desirable to enable querying of knowledge graphs through links curated by humans, and those predicted by AI.

Iain Marshall (King's College London) presented on RobotReviewer. Marshall spoke about RobotReviewerLive and his team's research to learn "whether we can use a machine learning system which is augmented with some human experts to try to keep systematic reviews up to date with low latency." With a database called Trialstreamer, his team is automatically collecting publications on randomized controlled trials daily. "We automatically extract structured data from them to identify the characteristics," Marshall explained. The tool is useful for maintaining and updating an existing SR using automated models and rules to screen and scan new trials, with the aim of identifying and collecting relevant studies for human review. Another project in an early stage piloted an SR of COVID vaccination candidates: the work identified 100% of the relevant studies, with a precision of about 40% (i.e., screeners included about 40% of the articles), he said. "It's quite a substantial improvement compared with conventional search," Marshall said. Another experimental path is using neural networks for generating narrative summaries automatically from trials. His team is exploring how to include humans in the loop to validate and edit these summaries to improve reliability.

Session moderator Joyce Tsuji (Exponent, Inc.) reflected that tolerance for AI errors is lower than for human error. She cited the example of car accidents caused by self-driving cars compared to those caused by human drivers. If AI tools can expedite regulatory assessments and fill important gaps by supporting health protective limits, what is the tolerance for errors? In this vein, Tsuji then asked about the opportunities for AI in supporting SR workflows and methodologies, and about the potential barriers to advancement. Jones noted that the HERO project has developed tools to help screen and focus on the likelihood of relevance, but opportunities remain to develop augmented human intelligence tools to tackle the most time-consuming aspect of science assessments: extracting information from articles. Watford, Lord, and Agai cited data extraction as a barrier.

There are also challenges with the PDF format. The vast majority of the literature is still unstructured data in the form of PDF, noted Agai, adding that researchers and academics are not under a lot of pressure to adapt and make the data extraction process more efficient. Tsuji agreed, highlighting the need for a breakthrough in how data are structured, to make them readily usable. Watford noted that, ideally, the basis for decisions would be provided along with the metadata published with the article or otherwise available in a structured repository of all of the supplemental information and data. Marshall noted that agreement on a shareable format is essential to solve the problem of data shareability. He added that collaborations between academia and industry might be helpful to promote standards of software development and usability. Watford added that standardizing units for this task can be extremely difficult and encouraged progress in this area.

Regarding pathology, Tsuji noted the challenges of interpretation by a machine and the benefits brought through the experience of a seasoned pathologist. At the same time, she asked if pathology's basis in pattern recognition provides an opportunity for toxicology or clinical studies. Watford mentioned a company called PathAI, which provides tools for machine-assisted pathology assessment, noting that it was not possible to exclude humans from the decision-making. Rooney added that NTP is actively pursuing AI technologies for histology because its histology slides contain an extensive amount of data. The project asks experts how much of the data they can actually use to reach reliable conclusions.

ENSURING RIGOR AND REPRODUCIBILITY IN ARTIFICIAL INTELLIGENCE APPLICATIONS

Marzyeh Ghassemi (Massachusetts Institute of Technology) discussed her work in designing machine learning processes for equitable health systems. She explained that when considering reproducibility, it is important to ask what it means for a model to perform well in different settings with different subgroups, and with different data. It is also important to understand the process that generated the model and provide caveats. Big-picture tools that accomplish this in a transparent way include data sheets for datasets and model cards for model reporting, she said.

Ghassemi said there are likely no simple fixes for ethical issues related to AI applications in health as well as other spaces. This is an ongoing process that needs diverse data and diverse teams. It also calls for considering the sources of bias in data, evaluating them comprehensively, and recognizing that not all gaps can be corrected.

Patel discussed probing the robustness of exposome phenotype associations with what is called multi-verse approaches. He said, "we can implicate many genetic variances all simultaneously, and say that they are robust and reproducible." He characterized the genome-wide association studies as a prime example of reproducible observational science, and that one area of improvement for the environmental health sciences is measuring exposures in human populations.

To explore the idea of reproducibility, Patel referenced the work of Steve Goodman and his colleagues at the Meta-Research Innovation Center at Stanford, who found that reproducible research is defined by its methods (is there enough detail to repeat the study?), its inferences (will users arrive at the same qualitative conclusion?), its results, and its robustness. Patel described robustness in terms of:

- How stable are results to variations and assumptions to the study design, modeling approaches, or for example, the seeds that are applied to stochastic algorithms for learning and for AI?
- The degree of determinism.
- Signal-to-measurement errors.
- Statistical criteria for validity claims.

Reif moderated the panel discussion, which was joined by John Absher (Squarespace, Inc.). Absher began by exploring the idea of rigor. He noted that SR requires "high precision and high recall," and one strategy is to "think about letting the machines do what they are good at—pulling out everything vaguely relevant—and then letting humans do that last mile where the precision becomes super important."

Ghassemi reflected on best practices for algorithm-assisted (machine) learning, including multiple data sources and results that can be reproduced and replicated. Reif highlighted the challenge in measuring data ranges widely across fields and applications, from tracking user behavior in an app to Patel's exposomics work, where "you are asking to measure everything forever for all people."

Ensuring rigor and reproducibility depends on solid source data, Reif said. Absher stressed the importance of collecting a representative sample that is as accurate as possible, because "the thing that haunts us all, no matter what context, are the things that we are not seeing," such as nonresponse bias.

Reif asked the panel about desired AI innovations for SRs. Patel said that his team is gathering summary statistics from papers and it is a struggle to extract data from the PDFs. Open datasets can inform how these summary statistics are produced and provide a way of filtering them for SRs. Absher added that attempting to extract data from PDFs is remarkably difficult, and that structured data are more valuable and useful than unstructured data. Ghassemi underscored the importance of public, open datasets that are heterogeneous, diverse, and representative.

Reif closed the panel by reflecting on the idea that data availability does not solve all problems, but it provides the information needed to address them. In the case of SR, it must be systematic, and it is not a system if it is not reproducible. Data are becoming available that can help address questions about rigor and reproducibility.

WORKSHOP SUMMARY AND FINAL THOUGHTS

In the final session of the workshop, members of the planning committee summarized topics that were explored during the workshop and shared examples of key takeaways (including opportunities) about the use of AI tools and open data practices in chemical hazard assessments.

Faustman highlighted the need for blended expertise (i.e., expertise in computer-assisted tools and relevant subject matter). She framed the concept of explainability as a challenge in the application of AI tools because some of the algorithms and search functions are not completely clear. Faustman also mentioned the prospect of expanding the application of ontologies and knowledge graphs as a way of acting on the taxonomization of environmental health entities for making SRs more efficient.

Auerbach reflected on the challenges that researchers face when using AI with SRs, especially the themes of reproducibility and transparency. Though AI tools can work quite well in filtering publications, it can be more challenging to extract data. He listed several potential steps: creating pre-structured data sources, systematic online living evidence summaries or repositories, metadata pre-annotation by authors and publishers, guidance for end users, curated resources, and open-source tools.

When considering how to optimize data extraction for evidence synthesis and decision-making, Wallace emphasized a need for more training data and labeled data. He mentioned the potential benefit of using alternative supervision strategies to extract data, such as a rule-based method, or distance supervision. He underscored the importance of transparency, which can help inform decision-making, because it enables an assessment of the provenance of the prediction or decision.

Reif indicated that, because bias can be amplified by the way in which input data are collected, it is important to inspect input data purposefully and carefully. He added that data representation and structure are key for all AI and machine learning applications. Reif indicated that open-source tools and data are expected to be important for promoting rigor, reproducibility, and public acceptance of the results.

Faustman asked the members to consider examples of key opportunities for AI to advance the SR workflow and methodology to enhance efficiencies. Auerbach underscored the use of crowdsourcing approaches as a theme related to the application of AI tools for SR. Faustman suggested the use of knowledge maps to indicate what different programs do for SR. Tsuji mentioned that it would be useful to compare all of the different AI tools on the basis of what they can do and how they go about it. That would help users decide what tool to select for a particular SR application. Wallace cautioned not to lose site of the importance of building prototype AI tools that can eventually become mature enough for common usage.

Tsuji outlined several barriers, including annotation and the need to go beyond the PDF to allow for data extraction in a more accurate and efficient manner. Wallace indicated that tools are available that expedite working with PDFs. Tsuji also mentioned that, because of the competitive environment they work in, some scientists can be reluctant to make all of their data available. Faustman mentioned possible funding incentives offered by agencies for sharing datasets. Tsuji underscored giving recognition to those who generated the data in the first place.

Wallace commented that a useful project would be to train an AI tool to convert PDFs in the open-access subset in PubMed Central into XML. He added that there is a sufficiently sized training dataset to do this.

As the conversation moved on to opportunities for implementation in the short and long term, Wallace highlighted the need for more annotated data that are readily available.

Nicholas Chartres (University of California, San Francisco) noted that tool performance will vary depending on the dataset. Although many tools are available, there is no one provider with a suite of tools that can be used throughout the SR process—which makes integrable tools invaluable, he noted. Chartres also emphasized the importance of validation being conducted by the researcher who implements AI tools within the SR process. Chartres added that the researcher can also explain why a given tool was chosen and the approach for attaining the results.

Regarding publications, Reif mentioned that there are many venues for research dissemination and it is important for SR methods and results to be tied to something citable. Faustman agreed and emphasized the need for code to be citable.

Chartres emphasized the need for original data to be made openly available so that data extraction does not become a bottleneck in the SR process. He also mentioned that standardized reporting is an important objective.

Wallace envisioned the next decade as an opportunity to be more audacious. Characterizing the traditional SR process as artificially constrained, he suggested the use of machine learning to monitor the entire pool of literature, instead of using Boolean queries to retrieve a subset of articles that are eligible for screening. Another idea he offered is to attempt to "curate and construct summaries of evidence on the fly," noting that the results would be imperfect.

The panel discussion ended after exploring various notions of bias in the context of computer-assisted data extraction for chemical hazard assessments.

DISCLAIMER: This Proceedings of a Workshop—in Brief was prepared by **JEANETTE BEEBE, KALEY BEINS, KATHRYN Z. GUYTON,** and **RAYMOND WASSEL** as a factual summary of what occurred at the workshop. The statements recorded here are those of the individual workshop participants and do not necessarily represent the views of all participants; the workshop planning committee; the workshop participants' institutions; or the National Academies of Sciences, Engineering, and Medicine.

REVIEWERS: To ensure that this Proceedings of a Workshop—in Brief meets institutional standards of quality and objectivity, it was reviewed in draft form by **DAVID REIF,** North Carolina State University, and **QINGSHENG WANG,** Texas A&M University. The review comments and draft manuscript remain confidential to protect the integrity of the process.

SPONSORS: This workshop was supported by the U.S. Environmental Protection Agency. Contract No. 68HERC19D0011.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2022. *Artificial Intelligence Tools and Open Data Practices for EPA Chemical Hazard Assessments: Proceedings of a Workshop—in Brief*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26540.

Division on Earth and Life Studies

NATIONAL ACADEMIES *Sciences Engineering Medicine*

The National Academies provide independent, trustworthy advice that advances solutions to society's most complex challenges.
www.nationalacademies.org